Methodological expectations for demonstration of health product effectiveness by observational studies

Michel Cucherat, Olivier Demarcq, Olivier Chassany, Claire Le-Jeunne, Isabelle Borget, Cecile Collignon, Vincent Diebolt, Marion Feuilly, Béatrice Fiquet, Clémence Leyrat, et al.

Therapie xxx (xxxx) xxx—xxx

Available online at ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

GIENS WORKSHOPS 2024 / *Clinical research and health product evaluation*

Methodological expectations for demonstration of health product effectiveness by observational studies☆

Michel Cucherat[a], Olivier Demarcq[b], Olivier Chassany[c], Claire Le Jeunne[d,e], Isabelle Borget[f,g,1], Cécile Collignon[h,1], Vincent Diebolt[i,1], Marion Feuilly[j,1], Béatrice Fiquet[k,1], Clémence Leyrat[l,1], Florian Naudet[m,n,1], Raphaël Porcher[o,p,1], Nathalie Schmidely[q,1], Tabassome Simon[r,1], Matthieu Roustit[s,*]

[a] *Metaevidence.org, service de pharmacologie, hospices civils de Lyon, 69000 Lyon, France*
[b] *Pfizer Inc, Chief Medical Affairs Organization, Pfizer US Commercial Division, 75014 Paris, France*
[c] *Unité de recherche clinique en économie de la santé (URC-ECO), hôpital Hôtel-Dieu, AP—HP, 75004 Paris, France*
[d] *Université Paris Cité, AP—HP, 75000 Paris, France*
[e] *Hôpital Cochin, 75014 Paris, France*
[f] *Gustave Roussy, Biostatistics and Epidemiology Office, université Paris-Saclay, 94810 Villejuif, France*
[g] *Inserm, université Paris-Saclay, CESP U1018, Oncostat, labeled Ligue contre le cancer, 94810 Villejuif, France*
[h] *Medtronic, 78620 L'Etang-la-Ville, France*
[i] *F-CRIN Coordination, 31300 Toulouse, France*
[j] *Bayer HealthCare SAS, département accès au marché, 59045 Lille, France*
[k] *Amgen SAS, 92400 Courbevoie, France*
[l] *Department of Medical Statistics, London School of Hygiene & Tropical Medicine, WC1E 7HT3 London, United Kingdom*

DOI of original article: https://doi.org/10.1016/j.therap.2024.10.052.

☆ The articles, analyses and proposals of the Giens Workshops are the sole responsibility of their authors and are without prejudice to the position of their supervisory body.

* Corresponding author. Centre d'investigation clinique, Inserm CIC1406, CHU Grenoble Alpes, 38043 Grenoble cedex, France.
E-mail address: MRoustit@chu-grenoble.fr (M. Roustit).

[1] The participants of the Round Table ''Clinical research and health product evaluation'' of Giens Workshops 2024.

https://doi.org/10.1016/j.therap.2024.10.062

Please cite this article as: M. Cucherat, O. Demarcq, O. Chassany et al., Methodological expectations for demonstration of health product effectiveness by observational studies, Therapies, https://doi.org/10.1016/j.therap.2024.10.062

[m] *University of Rennes, CHU Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) -UMR-S 1085, centre d'investigation clinique de Rennes (CIC1414), 35000 Rennes, France*
[n] *University Institute of France, 75000 Paris, France*
[o] *Université Paris Cité, université Sorbonne Paris Nord, Inserm, INRAe, Centre for Research in Epidemiology and Statistics (CRESS), hôpital Hôtel-Dieu, 75004 Paris, France*
[p] *Centre d'épidémiologie clinique, hôpital Hôtel-Dieu, AP—HP, 75000 Paris, France*
[q] *Takeda France, accès des patients à l'innovation, 75000 Paris, France*
[r] *Service de pharmacologie, plateforme de recherche clinique de l'Est parisien, Sorbonne université, AP—HP, 75012 Paris, France*
[s] *University Grenoble Alpes, Inserm, CIC1406, HP2 U1300, CHU Grenoble Alpes, 38043 Grenoble, France*

Received 2 october 2024; accepted 4 october 2024

KEYWORDS
Observational studies ;
Real world data ;
Real world evidence ;
Health technology assessment ;
Critical reading ;
Decision-making ;
Regulatory agencies ;
Market access

Summary The issue of assessing the effectiveness of health technologies (drugs, devices, etc.) through observational studies is becoming increasingly important as registration and market access agencies consider them in their evaluation process. In this context, observational studies must be able to provide real demonstrations of a level of reliability comparable to those produced by the conventional randomized controlled trial (RCT) approach. The objective of the roundtable was to establish the acceptability criteria for an observational study (non-randomized, non-interventional study) to be able to provide these demonstrations, and possibly serve as a confirmatory study for registration and market access authorities, the construction of therapeutic strategies or the development of recommendations. In order to do this, the study must be a real confirmatory study respecting the hypothetical-deductive approach and guaranteeing the absence of HARKing and p-hacking by attesting to the establishment of a protocol and a statistical analysis plan, recorded before any inferential analysis. It must also be part of a formalized approach to causal inference and demonstrate that it correctly identifies the causal estimand sought. The study should ensure that there is no residual confusion bias by taking into account all confounding factors affecting the comparison, which should be determined by a formal approach (such as a graphical causality approach, DAGs). Residual confusion bias diagnoses by forgery and nullification analysis should be non-existent. The study shall be at low risk of bias, in particular selection bias, among others by using a target test emulation design. Overall type I error risk should be strictly controlled. The absence of selective publication of results and selection bias should be ensured.

Abbreviations

ATE	average treatment effect among the treated
ATT	average treatment effect
AVK	antivitamins K
CFs	confounding factors
DAG	directed acyclic diagram
DOAs	direct oral anticoagulants
FDA	Food and Drug Administration
HARKing	hypothesizing after the results are known
HETE	hypothesis evaluating treatment effectiveness
IPTW	inverse probability of treatment weighting
ISPOR/ISPE	Society for Pharmacoeconomics and Outcomes Research/International Society for Pharmacoepidemiology
ITT	intent-to-treat
NUC	conditional exchangeability hypothesis
PI	principal investigator
RCTs	randomized controlled trials
RWD	real world data
SAP	statistical analysis plan
SMD	standardized mean difference

Introduction

Until now, observational studies were considered insufficiently reliable to be considered as pivotal studies in the initial assessment of the clinical benefit of treatments (in the broad sense of health technology) [1—3], in particular by the registration and market access agencies [4]. This is due to the presence of biases inherent in the observational approach (such as confusion, selection or classification biases), the absence of certain data in the sources used (missing data or data not collected), as well as other methodological issues such as HARKing, p hacking, statistical multiplicity, publication bias and selective reporting of results (*selective reporting)*. All these problems, and the lack of formal statistical language to express causal effects, did not allow the results of observational studies to be interpreted causally.

However, the registration and market access agencies have recently been considering, on a case-by-case basis, observational studies, such as studies on real world data in their evaluation [5—7]. This is particularly true for non-randomized non-interventional studies used as confirmatory studies (inferential studies) of the treatment effect in place of randomized controlled trials (RCTs) to inform regulatory, reimbursement and pricing decisions, and clinical care [8] through the construction of therapeutic strategies and medical recommendations.

This would not, however, represent a reduction in the requirement for evidence beyond a reasonable doubt to demonstrate the benefit of treatments, but would be justified by theoretical and methodological advances in epidemiology such as causal inference, emulation of target trials and many others [9]. These advances (or methodological innovations) suggest that, under certain conditions, it is possible to produce sufficiently reliable, high-level evidence-based results that can be used instead of or in addition to RCTs results, for decision-making, the definition or modification of therapeutic strategies, the development of recommendations, or the Health Technology Assessment.

Given the stakes involved in these decisions, the use of observational study results in place of RCT results requires high methodological requirements to ensure that these studies have a credibility close to that of RCTs.

Purpose of the roundtable

The objective of this round table was to define, based on the expertise of the participants, on the state of the art and the literature, the criteria of methodological validity so that the results of observational studies can be considered as evidence of the clinical benefit of a health technology (drug, device, etc.). These validity criteria thus define a list of expectations that evaluators and decision-makers might have to consider the results of such studies (Appendix 1). However, it does not prejudge the use that different agencies will make of this possible opening to observational studies confirming the treatment effect.

This list is also intended for the users and designers of these studies (knowing the methodological acceptability criteria to be met to maximize their chances of seeing the results of the study taken into account in a decision-making process). More generally, it makes it possible to know the conditions making it possible to conduct a quality causal observational study.

Table 1 Hypothetical use case examples (without prejudging their actual acceptability by agencies).

- A new-generation cholesterol lowering agent placed on the market based on its pharmacological effect (LDL decrease) and whose clinical benefit on cardiovascular mortality would be demonstrated by an observational study
- A left ventricular assist pump placed on the market based on performance criteria and whose impact on mortality in cardiogenic shock would be demonstrated later by an observational study
- Demonstration of superiority in terms of glycemic control of one basal insulin over another due to its improved pharmacokinetic properties
- Gene therapy for hemophilia placed on the market based on a non-comparative study and the superiority of which to the supplementary treatment would be obtained subsequently by an observational study

Terms and scope of the roundtable

The field of observational studies is very broad, bringing together studies of a very wide variety by: their objective (descriptive or analytical, ranging from epidemiology and public health to health technology assessment or interventions), their ambition (exploratory or confirmatory studies), their design (cohort studies, case-control studies, cross-sectional studies, self-controlled studies, etc.), the data sources used (primary data, administrative databases, electronic medical records, disease registers, treatment registers, etc.), and their method of analysis (from conventional multivariable regression to more advanced methods, including machine learning).

In view of the objective of the roundtable, the discussion focused exclusively on comparative studies assessing the clinical benefit and safety of a health technology or intervention and whose purpose is to change the therapeutic strategy (regulation, recommendations, market access, etc.).

The discussion of the roundtable did not concern studies using an external control group, which was the subject of a previous roundtable in 2019 [10].

Although some of these approaches have been validated empirically [11], no examples of decision-making exist at present (as of June 2024) given the recent evolution of the agencies' positions. However, this use is often claimed for different scenarios, some hypothetical examples of which are given in Table 1.

Why acceptability criteria and not methodological recommendations

The Working Group focused on defining the methodological validity criteria to be met by the results of confirmatory

observational studies, not on making recommendations on how to design or carry out such studies (which are already available) [8,12—15]. Indeed, given the issue at stake, the question is the actual reliability of the results obtained and not the quality of the study design.

For example, for confusion bias, the challenge is to obtain results without any residual confusion bias (or with negligible residual confusion bias) and not ''only'' to implement a confusion bias control approach. Even a correct confounding bias correction approach may fail to achieve its objective for many reasons (unmeasured confounding factors, statistical model not adapted to the data, etc.).

In this context, there is a real obligation to achieve results (a negligible residual confusion bias) and not simply an obligation to use means (a process of correcting the confusion bias). Indeed, with the observational approach, there is no guarantee that the methods used to correct biases will produce the desired effect, because everything depends on the adequacy of the data with the assumptions made by these methods. The observational approach can only isolate the treatment-specific effect if the validity assumptions associated with the observational approach (causal inference assumptions) and the analytical methods used are effectively verified by the data. The justification of the plausibility of these hypotheses is therefore crucial (identification analysis) [9].

Unlike RCTs, which avoid most design biases (and whose correct, undistorted design ensures the reliability of the results), in observational studies, biases are present at the source. This is then corrected by analysis, but success depends on the extent to which the data verify the assumptions of validity of the methods used.

The evaluation of these studies is therefore not simply a matter of verifying whether the methodological principles have been correctly used (as with the RCTs), but rather of ensuring that the validity assumptions of the methods are verified in the context of carrying out the study and that the general principles of the scientific approach, such as the hypothetical-deductive approach, have been implemented.

Since this certainty is not easy to obtain, when this is not completely ignored [16], investigators sometimes make statements such as: *''results are likely to be affected by confounding bias and should be interpreted with caution''*, *''However, because observational studies are prone to confounding and selection bias, causality cannot be affirmed''*, *''Because claim databases can be vulnerable to selection and confounding bias, these results are statistical associations but not causal''*.

With the objective of producing evidence of clinical benefit, this type of reservation makes usability of the results obsolete. Only works where it is possible, properly, to refrain from mentioning such limits and from concluding that there is causality will be admissible for this use (Box 1) [17].

Box 1 : Studies on ''real-life data'', a name to be avoided.

The term ''real world data'' (RWD) studies is commonly used, primarily to distinguish them from randomized trials. This term is ill-suited, since the problem of these studies is not really the nature of the data, but rather the observational character which determines their methodological limitations [17]. It is the methodology and robustness of the results that fundamentally oppose randomized trials and observational studies, not the conditions for data generation. The term ''real-life data'' is also inappropriate, because all studies are based on real data. At this level, it is rather proposed to speak of data collected in current practice.

Methodological acceptability criteria for confirmatory observational studies

In the current state of knowledge, it is possible to identify the conditions of validity of results from observational studies of treatment effect inference [18].

Acceptability criteria related to the basic principles of the scientific approach

Confirmatory study

In order to comply with the prevailing hypothetical-deductive approach to treatment benefit assessment (see ICH E9), the observational study should be a confirmatory study, specially conducted to test a pre-specified hypothesis, established prior to the analysis of the data [19—21] These confirmatory studies are referred to as the hypothesis evaluating treatment effectiveness (HETE) study in Society for Pharmacoeconomics and Outcomes Research/International Society for Pharmacoepidemiology (ISPOR/ISPE) best practices [22].

Guarantee of absence of HARKing

A major limitation of retrospective studies, i.e. studies for which the data source already exists at the time of the formulation of the study objective and its planning [23], is the possibility of HARKing. HARKing (hypothesizing after the results are known) occurs when the hypothesis is formulated from a preliminary analysis of the data used for the study itself. The study is no longer an independent confirmation or refutation of the hypothesis. In the case of HARKing, the results of the study cannot go against the hypothesis and the reasoning is completely tautological: the hypothesis is not precondition to the study, but is directly derived from the observed results, which in turn confirm it. The approach therefore does not follow the hypothetical-deductive approach and remains inductive because the hypothesis is not tested but is constructed from the data.

In retrospective observational studies, ensuring the absence of HARKing is not simple and involves at least recording the protocol before analysis [21], or even access to data subject to the presentation of a protocol, or traceability of the chronology of access to databases. However, none of these means is sufficient (in particular where the investigators also manage the data as in the registers). As suggested by the good practices of ISPOR/ISPE [20], an explicit commitment by the authors is ultimately expected that the hypothesis was generated independently of the data set used for the study (in the protocol, the SAP statistical analysis plan, the analysis report and the publication).

No p-hacking guarantee

P-hacking involves adapting the analysis of the data to the results produced and publishing only the results in favor of the hypothesis tested (''torture the data long enough and they will confess to anything'', Ronald H. Coase). This practice invalidates the results produced, since it has been shown on numerous occasions that it is feasible to produce, with the same set of data, a large number of different results [23]. Meta-epidemiological studies have shown the frequency of this practice [24,25], which may be legitimate in exploratory studies but prohibitive for confirmatory studies [17].

To rule out the possibility of p-hacking, a protocol and a statistical analysis plan are expected to be drawn up before any inferential analysis of the data [24], recorded [25—27] with documentation of the chronology of access to the data and an attestation by the authors that the inferential analysis was carried out in accordance with the statistical analysis plan drawn up a priori [27].

In particular, it is necessary to separate the analyzes of 'preparatory' data intended to qualify the data source(s) from the inferential analysis on the judgment criterion(s) [24,28].

Recommandation 1

It is recommended to consider for the decision-making process only studies that are part of the confirmatory hypothetico-deductive approach and for which HARking and p hacking can be ruled out due, for example, to the guarantee that the objective, the protocol and the statistical analysis plan have been established prior to any inferential analysis and that this analysis has been carried out in accordance with this plan; a guarantee provided by the attestation of the authors, the registration or publication of the protocol and SAP, etc.

Food and Drug Administration (FDA guideline lines [8] where same concepts are mentioned): lines 51, 124, 112, 90.

Acceptability criteria for demonstrating causality

Using a causal inference approach

Causal inference theory [29—32] identified that a causal conclusion from an observational (and thus non-experimental, non-randomized) comparison was possible provided that several fundamental assumptions were tested by the data and how to analyze them:

- the positivity hypothesis: all patients in the target population can receive all the treatments compared. For example, this assumption makes it impractical to evaluate a new treatment from routine care data that would be routinely used after it is available in all patients. In this context, the probability that a new patient will receive the comparator treatment will be virtually nil, and the positivity hypothesis will not be respected;
- the consistency assumption: for a given patient, the fate observed with a specific treatment corresponds to the potential fate under this same procedure. In other words, this means that patients in a group have received a version of treatment that fits the definition of treatment assigned to that group (as in a randomized trial where all patients in a group are treated according to the same protocol);
- the non-interference hypothesis: the treatment of some patients does not affect the potential fate of other patients. This hypothesis can be challenged with vaccines, for example, because vaccine coverage protects unvaccinated patients. In this situation, a comparison between vaccinated and non-vaccinated could conclude that the vaccine is not of interest;
- the conditional exchangeability hypothesis (NUC hypothesis: no uncontrolled confounding): it means that the level of risk (frequency of judgment criterion) of the comparator group, if it had received treatment, would be the same as that of the treated group. In other words, the validity of this hypothesis corresponds to the absence of confusion bias.

In causal inference, the validity of all these hypotheses, which are not all testable on the data (such as the NUC hypothesis), must be verified. The plausibility of these hypotheses is then established by means of sensitivity analysis and by understanding the clinical context. If this is not the case, it is referred to as ''identification bias'' [9,32]. At the decision-making level, the expectation is a strong justification that these assumptions are verified by the data used and how they are analyzed. In the properly designed and conducted randomized trial no hypothesis is needed to conclude causally.

Counterfactual reasoning

Causal inference requires counterfactual reasoning (based on the concept of ''potential outcome'') which requires knowing for each patient the potential fate under treatment and without treatment. But in practice, only one of

Table 2 Some possible estimands in terms of target population for ''treatment effect'' in observational studies.

Name	Definition	Corresponding/Target Population	Causal Question
Average treatment effect (ATE)	The ATE can be interpreted as the difference between the outcome that would be observed if all subjects were treated, versus untreated; this value is therefore not calculable (these are ''potential endpoints''), and must be estimated from the results observed in both treated and untreated patients.	The entire population	Can treatment be recommended to all future patients (comparable to those eligible for study)?
ATT (ATE among the treated)	Corresponds to ATE in treated patients only	Patient population with characteristics similar to those treated (i.e. with the study treatment)	Confirm that the treatment is beneficial to those currently being treated with it?
ATC (ATE among the controls)	Corresponds to ATE in patients in the control group only	Control patient population	Can treatment use be extended to those who are not currently receiving it?

these two ''potential outcomes'' (the other being counterfeit) is observed. Causal inference theory provides methods for estimating counterfeit from available patient information. The design of the study must therefore be comparative: cohort study, case control study or self-controlled design, time series interrupted, difference of differences, etc.

Purely descriptive studies do not allow for causal inference and are inappropriate, such as simple ''*change from baseline* '' comparisons that do not isolate the treatment effect in an observational approach [33].

Estimand

In observational studies, there are several ways of conceiving the treatment effect [34—36]. Each of these ways corresponds to a different estimand.

> *However, in this sense, the estimand does not correspond exactly to the same definition as that used in the standard randomized trial [37], where it has a more general meaning compared to the basic concept which designates the conceptual quantity that one wants to apprehend in a statistical analysis/a causality search [38].*

The different possible estimands in an observational study are listed in Table 2. Each estimand involves an appropriate method of analysis: for example, a 1:1 pairing among the treated group gives an average treatment effect in the treated population (*average treatment effect among the treated* [ATT]), while the *inverse probability of treatment weighting* - IPTW [with appropriate weights]) gives the average effect in the population representing all study participants (*average treatment effect* [ATE]).

For confirmatory studies for the evaluation of a health technology, the expected estimand is rather ATE (in the total study population), since it corresponds directly to the direction of treatment effect conventionally measured in the randomized trial and is the natural estimand in a target trial emulation.

Recommandation 2

It is recommended to consider for the decision-making process only studies that:

- satisfactorily implementing an inference approach with a clearly causal issue;
- based on counterfactual reasoning due to comparative design (based on emulation of a target trial, see below);
- and with an estimand directly corresponding to the causal question and satisfying the assumptions of causal inference (identifiability)

FDA: 138, 37, 223.

Intention to treat analysis/treatment received

In terms of estimand [37], the treatment effect can be viewed in two different ways: the effect of the decision to treat (intention to treat) or the effect of actually taking treatment (treatment received). The first seeks to understand what can be expected from treatment in terms of the a priori chance of changing the patient's fate, when the doctor makes the decision to use this treatment and whatever will happen during treatment (compliance, adverse effects, competitive risks, etc.). It corresponds to the classical analysis in intent to treat. After determining treatment initiation, the analysis will consider all events occurring during the pre-defined follow-up period (e.g. 3

years) and including those occurring after the patient has stopped treatment initiated if the discontinuation occurs before the end of the follow-up.

The treatment analysis received will only consider periods of exposure, during which the patient receives the treatment in question (studied or comparative). Follow-up stops when treatment is stopped. This approach therefore seeks to determine what treatment will bring to observant patients, who tolerate it, and who will not experience an intercurrent event, a particular situation that cannot be anticipated at the time of the decision to use a given treatment for a patient.

In confirmatory health technology assessment studies, an intent-to-treat (ITT) approach (for superiority, as is the case in the usual pivotal randomized trial) is expected, which corresponds to a treatment policy estimate.

Recommandation 3

It is recommended that only those results that correspond to an estimand appropriate to the decision and its causal question, typically the ATE estimand, should be considered, which leads to a preference for inverse- treatment- probability weighting (with the right weights) or multivariate regression followed by standardization (g-computation) over matching methods.

Similarly, it is recommended that only the results of an ITT analysis be considered for the decision, with a treatment policy for the management of intercurrent events.

Acceptability criteria for confusion bias

Consideration of all confounding factors

Confusion ''bias'' is a major problem in observational studies due to the lack of randomization. Selection bias aside, confusion bias is certainly responsible for most of the inconsistencies in the conclusions between observational studies and randomized trials.

The purpose of the analysis is to correct the result of this bias. However, this can only be achieved, at a minimum, if all the variables allowing the control of this bias are included in the analysis, either in the form of a direct adjustment (multivariable regression) or by methods of the weighting or standardization type.

The expectation is therefore clear: an analysis considering all the confounding factors (CFs) of the study. But judging this is not easy, as the identification of the CFs is not trivial, and the reader is not necessarily able to determine them by himself in order to verify the completeness of the list used by the authors.

The list of CFs considered will therefore have to be justified in order to provide this guarantee. This justification should be based on a formalized approach:

- identification of variables associated with endpoints from previous knowledge, ideally through systematic literature review to ensure completeness. Expert opinion alone, and even less the study data itself, cannot meet this requirement;

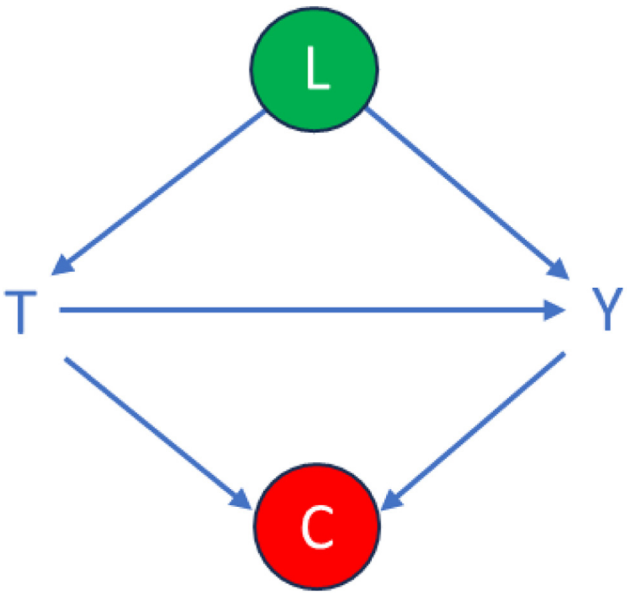

Figure 1. Example of a directed acyclic diagram (DAG), which represents the causal relationships between the various variables. An arrow represents a causal (potential) effect, and an absence of an arrow represents the absence of any causal relationship between two variables. The graph is said to be directed because all links are oriented, and acyclic because no path should form a closed loop. In this example, T represents the treatment, and Y represents the patient's fate (outcome). L is a confounding factor, which should be considered in the analysis. It is a collider on the T → C ← Y path that blocks this path and should not be considered in the analysis.

- the identification of all interrelationships between these variables and between these variables and the treatment studied, using knowledge of practices;
- the formalization of these dependencies by an directed acyclic diagram (DAGs) [39,40] (Fig. 1).
- the identification of a *minimum sufficient adjustment set to* block all paths leading to a confusion bias.

This formalization makes it possible to establish the list of covariates that must be considered in the conditioned analysis in order to eliminate the confusion bias. It also identifies covariates not to be used to avoid introducing other ''biases'' such as colliders or mediators [41—43].

In order to verify whether this expectation is met, it is necessary to verify the rigor of the procedure for identifying the covariates to be taken into account and to verify that these variables were properly measured and effectively taken into account in an appropriate manner.

Beyond this preliminary approach, it is also necessary that the method of analysis has fulfilled its role correctly, because this is not guaranteed to be 100%. For example, with a method of ''matching'' by the propensity score, it is not 100% guaranteed that the CFs considered are distributed in the same way in the 2 groups, which is nevertheless the objective.

It is therefore necessary to show the correct performance of the analytical method. The elements needed to document this vary depending on the method used: overlap of distributions of propensity scores between the 2 groups, comparison of distributions of covariates after pairing or weighting or measurement of the imbalance of each covariate (the p

value is inappropriate in this situation, an standardized mean difference (SMD) is more appropriate even if it has certain limitations), etc.

There is a wide variety of validated statistical methods that can be used to correct the confusion bias. All these methods work well when the causal inference model used matches the data and is therefore equivalent to a few details. And there is no magic method, ensuring the elimination of any confusion bias for sure (even with AI methods). Therefore, it is futile to try to judge the truth of the result through the method used. For this reason, diagnostics of the reliability of the results produced have been developed.

A demonstration of the absence of residual confusion bias

Even with an optimal approach to identification and the use of an analytical method to account for CFs, there is always a possibility of residual confusion bias (due to an unidentified factor or a poorly adapted/specified statistical model or other reasons).

Such a residual bias must be capable of being ruled out by the results obtained at the level of negative [44—47] (or positive controls if the result is an absence of association) or by those of the quantitative bias analysis [48] (robustness analyzes more specific than the usual sensitivity analyzes, specially designed to test the robustness of the result obtained with respect to the confusion bias, such as the E value [49,50]).

Recommandation 4

It is recommended to consider for the decision-making process only studies for which:

- a formalized approach (DAGs) was used to identify the CFs affecting the study;
- all the CFs affecting the study could be taken into account without over-adjustment;
- the method of considering the CFs fulfilled its role correctly (for example, for a matching method, correctly rebalance (SMD < 0.10 for example) the distribution of the CFs between the 2 groups, etc.);
- residual confusion bias is shown to be negligible using negative (or positive) controls or a well-conducted quantitative bias analysis.

FDA: 76, 80, 150, 139, 188, 226, 229.

Acceptability criteria for other biases

Ultimately, a low to moderate risk of bias

Observational studies are exposed to other biases in addition to the confusion bias. These other biases are equally likely to distort the results. They can be grouped into 5 categories: classification, deviations from intended interventions, missing data, measurement and selective reporting bias. It is not possible to develop all of these biases in this report, but there is a common bias risk assessment tool for these observational intervention studies, ROBINS-I [51].

Using this tool, it is possible to assess the risk of bias in a study, which the evaluator expects to be low or moderate.

Data quality

The issue of data quality is crucial but complex [24,28].

Data may be biased due to poor classification of exposure or measurement errors in judgment criteria. The error can be symmetric (affecting both groups in the same way) or asymmetric (the error depends on the treatment received). The consequences of these two types of errors are different depending on the nature of the result. Symmetrical errors cannot induce bias toward incorrectly concluding a difference, but they can bias toward *bias toward the null* and can challenge a conclusion about the absence of difference, for example in a non-inferiority study. Asymmetric errors can lead to bias regardless of the nature of the result.

> *The measurement error should not be seen as a simple problem of recording of the occurrence of the endpoint in the database, as the mechanism of bias is sometimes complex. For example, to compare the frequency of major bleeding between direct oral anticoagulants (DOAs) and conventional antivitamines K (AVKs), a study using claims databases takes hospitalization for hemorrhage as a judgment criterion. This study is carried out rapidly after the marketing of the DOA. Asymmetric error in the ''measurement'' of the judgment criterion is to be worried about, as physicians will have little practical experience with DOAs unlike AVK. It is likely that hospitalization will be more common with bleeding of the same intensity in DOA patients than in AVK patients, especially since no marker comparable to INR is available for DOA. As a precaution, doctors will prefer to hospitalize patients on DOA. The error therefore occurs at the root, in the process leading to the measurement, i.e. hospitalization. The chain of measurement has different performances between the 2 groups, with more false positives in the treated group. Furthermore, the problem will not be detected by validating the correctness of the reason for hospitalization.*

Measurement errors on the CFs may render the conditioned analysis obsolete, invalidating the possibility of correcting the results of the confusion bias.

On this issue of data quality, validation of the data and algorithms used in terms of accuracy and completeness (missing data) is expected. This validation shall be based on a systematic check of the positives and negatives (exposure and judgment criteria) or possibly on a sample [28]. These validation studies will also identify the assumptions required for quantitative baseline analyzes to assess the robustness of results against data biases [48,52,53].

Another approach to data validation (which also covers the validation of CFs consideration) is the use of positive controls, i.e. to show that known results (those of previous randomized trials in the field) are found by analyzing these data. The absence of p-hacking in this validation is essential to convince. This validation can be carried out globally for the data source (by its manager) prior to its exploitation by studies in the form of a prior data validation (''benchmarking'') [54,55], or specifically study by study (carried out by the study authors) [56].

There is also a need to adapt existing data sources (particularly registers) to make them usable for this standard of

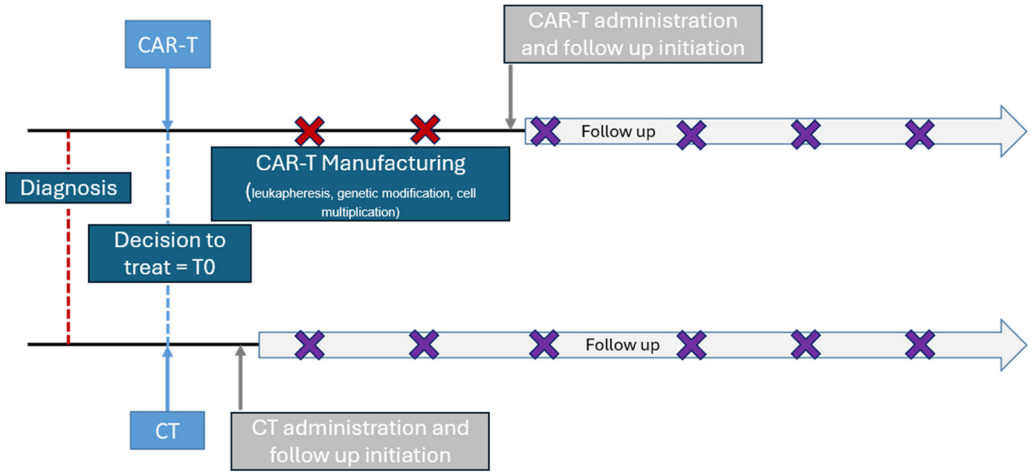

Figure 2. Example of an immortal time bias selection bias. With T-cells, there is a delay between the decision to use this treatment and its actual injection (the time to carry out leukapheresis and the transformation of the T-lymphocytes of the patient before re-injecting him). During this period, deaths occur (represented by red crosses). In a database study, the only way to identify patients treated with CAR T-cell may be to mention the injection. These patients are then compared to patients treated conventionally, by chemotherapy (CT), and who are identified by the fact that they have received this chemotherapy. A selection bias of the type of immortal time bias then affects this study, because in the CAR T-cells group, patients who are eligible, because the decision to use this treatment has been taken, will not be included because of their death before injection (red cross). In contrast, early deaths in the CT group are recorded and counted (purple crosses). Without any difference between the two treatments, there will necessarily be more deaths in the CT group than in the CAR T-cells group. Patients are therefore not included because they are in the CAR T-cells group and have presented the judgment criterion (death), the 2 conditions that lead to immortal time bias. For an eligible patient, for whom it was decided to treat with CAR T-cells, early death leads to non-inclusion, whereas for a similar patient for whom it was decided to use chemotherapy, death occurring on the same date after the decision to treat does not lead to non-inclusion. To avoid this selection bias, the groups should be defined by the intent to use CAR T-cells or CT treatment, and follow-up should begin at the time of this decision, which will synchronize, in the natural history of the disease, the starting t0 of follow-up (and event counts) between the 2 groups. This will only be possible if the database contains, for example, the minutes of the multidisciplinary meeting that makes the decision to deal with it. However, there are statistical methods to account for this bias (e.g. clone-censor-weight approach).

observational studies, if such exploitation is desired by the data source managers. This includes including the judgment criteria necessary for the evaluation of treatments, including all potential CFs from these studies (which could also involve harmonization and publication of DAGs by clinical situation). Finally, managers of these databases should also anticipate variables that can be used as negative or positive controls and anticipate data validation.

Recommandation 5

Adapt data sources to become usable to meet these acceptability criteria for confirmatory observational studies by ensuring that there are judgment criteria necessary to evaluate treatments, confounding factors and variables that can be used as negative or positive controls.

Selection bias

Selection bias is a major issue in observational studies, and certainly as important as confusion bias. This complex bias occurs when some eligible patients (or follow-up periods) are not included for a reason that depends on both treatment and judgment.

The term ''selection bias'' is misleading, as it is not a representativeness issue (e.g., selection only of ECOG 0 patients) nor does it correspond to its synonym used in the context of randomized trials. This term is also sometimes used to refer to confusion bias.

It corresponds to the good synchronization of the start of monitoring (t0) between the two groups, which is typically randomization in RCTs. This figure illustrated in Fig. 2.

Recommandation 6

It is recommended that only studies with low or moderate overall risk of bias assessed by ROBINS-I be considered for decision-making.

In particular with a low risk of selection bias due, for example, to the good synchronization of the t0 of the monitoring between the two groups, and a low risk of classification and measurement bias evidenced, for example, by data quality validation studies.

FDA: 77, 139, 165, 167.

Target trial emulation

The emulation of a target trial is to mimic, when conducting the observational study, what would happen if a randomized trial had been conducted to address the same research question [31,57—59]. This translation of the randomized trial framework into the observational study framework provides

a better understanding of and fix to a number of observational study bias and design defects [59].

With the previous example of CAR T-cells, the problem of follow-up lag becomes obvious when one emulates what happens in a randomized trial. Follow-up begins in both groups at randomization. Deaths occurring in the CAR-T group prior to injection will be well considered in the analysis of the trial.

This emulation framework also aligns the causal issues of the observational study with those of the efficacy and safety of treatments as addressed in confirmatory trials.

Observational studies are traditionally attributed the role of evaluating the ''effectiveness'' of treatments while the trial estimates ''efficacy''. ''Efficacy'' is defined [23] as the effect of treatment under experimental conditions while ''effectiveness'' is the effect under routine conditions of use. This ''effectiveness'' is more akin to a public health question: what the impact of treatment in practice is given the profile of patients actually treated, actual compliance outside the experimental setting, misuse or off-label uses [60]. It should be noted that the FDA uses the term ''effectiveness'' in a different sense, which is more of a clinical benefit [61].

The function of these observational decision-making studies, like the pivotal randomized trial, is to demonstrate that the treatment provides a clinically relevant benefit, without unnecessary selection of the patients involved, and not to assess the impact of the practical use of this treatment (which is also of real interest, but this is another question). In particular, emulation of a clinical trial aligns the conduct of the inferential observational study with this objective.

The emulation process involves first developing a synopsis of a randomized trial protocol, and then constructing the observational study protocol to emulate each item of the protocol.

The emulation of a target trial has been empirically evaluated in the DUPLICATE-RCT project, [62] which shows that observational studies carried out using this approach can only find the results of randomized trials emulated in 75% of cases and with an effect correlation of 0.82. There are therefore discrepancies and work is under way to identify the factors leading to these discrepancies [63]. In any case, although emulation alone does not automatically guarantee the reliability of the results, it is an important contributing factor and its use is becoming increasingly essential.

On this point, a justification that the emulation was satisfactorily obtained, in accordance with the PRINCIPLED practical guide, will be expected, including the presentation of the test synopsis and a 'third column' describing the emulation of each point (see Table 1 of ref [24]). Particular attention will be paid to these aspects, since currently meta-epidemiology reveals that most claims of using an emulation approach are in fact abusive [64,65].

Recommandation 7

It is recommended that only studies for which an emulation approach of a target trial has been used and correctly performed should be considered for the decision, for example, the presentation of the protocol or synopsis of the emulated randomized trial, and satisfactory emulation of all points of that protocol and appropriate statistical analysis with respect to the design of the study.

FDA: 138.

Acceptability criteria related to other aspects

Strict control of overall type I error risk

The multiplicity of statistical comparisons due to multiple judgment criteria, comparative treatments or populations and subgroups is very common in observational studies [66]. As with randomized trials, this multiplicity causes overall type I error risk inflation leading to a significant risk of incorrectly concluding that there are no differences. Therefore, strict control of overall type I error risk is expected using the same methods as those used for the randomized trial, such as co-primary endpoints or prioritization, with or without type I error risk reallocation.

The results which can be used to infer an effect of the treatment studied must therefore be statistically significant in terms of overall alpha risk and not simply nominally significant.

Intermediate analyzes are sometimes performed as the base is populated (especially in a patient registry study of an infrequent pathology). Because of the multiplicity they cause, these intermediate analyzes must be anticipated and carried out with appropriate statistical methodology (e.g. Haybittle-Peto, O'Brien and Fleming).

Recommandation 8

It is recommended that only results for which the overall type I error risk is fully controlled should be taken into account in the decision, for example by using a standard method for managing the multiplicity of statistical comparisons (prioritization, allocation, reallocation, management of interim analyzes).

FDA: 242.

Informativity of study reports allowing critical reading

The informativeness of a study's report is essential to be able to judge whether it meets all these criteria for methodological acceptability. In the absence of recommendations specific to confirmatory observational studies, it is therefore expected that the report will follow the recommendations of current drafts of observational studies such as STROBE [23], RECORD-PE [67] or STaRT-RWE [68] by adding the description of the emulated target test (reported following CONSORT pending a specific guidance being prepared).

Publication bias

For the same treatment comparison, at a given time, it is generally possible to carry out several similar observational studies using the different data sources available. This multiplicity can be the anchor of a significant publication bias if only the ''positive results'' are published (or presented to agencies).

To rule out this possibility, a study report is expected to specify the number of similar studies undertaken by the same sponsor or principal investigator (PI) and the results of each of these other studies (see multi-base studies below). Although the registration of retrospective study protocols has many limitations, it is a lever for attempting to prevent and detect publication bias. It is undoubtedly important that this a priori recording details things a little different from the RCTs and in particular the previous knowledge of the base. A systematic search of other studies should be undertaken when preparing to use a particular study for decision-making purposes.

Another approach to the issue of publication bias is the conduct of multibase studies. While not a perfect bulwark, this approach also allows for the documentation of the stability or non-stability of results across data sources from the outset and could lead to better reproducibility of results [11,69].

Recommandation 9

It is recommended that only studies for which it is possible to rule out that they have been selected (and promoted or presented in a dossier) on the basis of their results should be considered for the decision-making, for example because of an attestation by the sponsor or lead investigator that no other similar studies have been carried out for the same purpose, that a systematic review does not find any evidence of other published or registered studies, or that the study is based from the outset on several data sources (multi-base studies).

If more than one similar study is available, the decision should consider the appropriate synthesis of these studies and not a single study in isolation.

Conclusion

This work has deliberately chosen to focus on the proposal of recommendations to assist evaluators or decision makers, whether health authorities or prescribers, in assessing the quality and reliability of the results of confirmatory studies proposed by study sponsors, rather than generating recommendations on the design of these same studies. Obviously, and a fortiori if the health authorities decide to appropriate this work to make these recommendations their expectations, they can inspire the promoters in their role of designer to develop the protocols and conduct and report the confirmatory studies to respect the validity/reliability criteria developed in this document.

This analysis suggests that, to meet the methodological acceptability criteria needed to conclude for decision-making purposes, observational studies will have to implement sophisticated approaches that are still little known and little used outside specialized circles.

Real progress is needed, because, with few exceptions, the studies currently being carried out to show the value of health technologies are far from meeting all these criteria of methodological acceptability and thus produce results that are sufficiently reliable to induce a change in the therapeutic strategy or to be considered in the decision-making process of health authorities.

On the other hand, to accompany this transformation, it will be necessary to raise awareness of and train in these new approaches and requirements, both the prescribers of these studies and the readers and decision-makers.

Recommandation 10

To raise awareness of the differences in the finality, objectives and methodology of these confirmatory observational studies compared to ''conventional'' studies, and to train in the concepts of causal inference and modern epidemiology.

The list of these acceptability criteria was established based on current knowledge and experience.

There is no guarantee that, in the state of knowledge, it will be possible to identify all the situations that may arise in practice. Even if all these criteria for acceptability are met, there is no guarantee that the results are conducive to decision-making use. On a case-by-case basis, these criteria may appear to be insufficient depending on the situation, particularly in the case of small numbers.

However, while these expectations are not being met at this time, recent methodological advances have already had a visible impact on the quality of published observational research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.therap.2024.10.062.

Disclosure of interest

The authors declare that they have no competing interest.

References

[1] Fanaroff AC, Califf RM, Harrington RA, Granger CB, McMurray JJV, Patel MR, et al. Randomized trials versus common sense and clinical observation: JACC review topic of the week. J Am Coll Cardiol 2020;76:580—9.
[2] Fonarow GC. Randomization - there is no substitute. JAMA Cardiol 2016;1:633—5.
[3] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely

collected data and subsequent randomized trials: meta-epidemiological survey. BMJ 2016;352:i493.
[4] Wieseler B, Neyt M, Kaiser T, Hulstaert F, Windeler J. Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy? BMJ 2023;380:e073100.
[5] Concato J, Corrigan-Curay J. Real-world evidence — where are we now? New Engl J Med 2022;386:1680—2.
[6] European Medicines Agency. Real world evidence framework to support EU regulatory decision making; 2023 [Accessed 7 October 2024 (88 pp.)]) https://www.ema.europa.eu/en/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained-regulator-led-studies-september-2021-february-2023_en.pdf.
[7] Burns L, Le Roux N, Kalesnik-Orszulak R, Christian J, Hukkelhoven M, Rockhold F, et al. Real-world evidence for regulatory decision-making: guidance from around the world. Clinical Ther 2022;44:420—37.
[8] Center for Drug Evaluation, Research (CDER), Center for Biologics Evaluation and Research (CBER), Oncology Center of Excellence (OCE). Real-world evidence: considerations regarding non-interventional studies for drug and biological products guidance for industry; 2024. https://www.fda.gov/media/177128/download. [Accessed 7 October 2024].
[9] Dahabreh IJ, Bibbins-Domingo K. Causal inference about the effects of interventions from observational studies in medical journals. JAMA 2024;331:1845—53.
[10] Cucherat M, Laporte S, Delaitre O, Behier JM, et al., participants à la table ronde « Recherche clinique » des Ateliers de Giens XXXV. Des études mono-bras aux études de comparaison externe. Considérations méthodologiques et recommandations. Therapie 2020;75:13—9.
[11] Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB, et al. Transparency and reproducibility of observational cohort studies using large healthcare databases. Clin Pharmacol Ther 2016;99:325—32.
[12] Desai RJ, Wang SV, Sreedhara SK, Zabotka L, Khosrow-Khavar F, Nelson JC, et al. A PRocess guide for INferential studies using healthcare data from routine ClinIcal Practice to EvaLuate causal Effects of Drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. BMJ 2024;384:e076460.
[13] European Medicines Agency. Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence - Scientific guideline; 2024. https://www.ema.europa.eu/en/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence-scientific-guideline. [Accessed 7 October 2024].
[14] CIOMS. Real-world data and real-world evidence in regulatory decision making: CIOMS 2024. https://cioms.ch/publications/product/real-world-data-and-real-world-evidence-in-regulatory-decision-making/#:~:text=This%20report%20was%20developed%20to%20inform%20discussions%20about%20the%20use. [Accessed 7 October 2024].
[15] Haute autorité de santé. Études en vie réelle pour l'évaluation des médicaments et dispositifs médicaux; 2021 [Accessed 7 October 2024] https://www.has-sante.fr/jcms/.
[16] Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. J Clin Epidemiol 2018;93:94—102.
[17] Pacheco RL, Martimbianco ALC, Riera R. Let's end ''real-world evidence'' terminology usage: a study should be identified by its design. J Clin Epidemiol 2022;142:249—51.
[18] Franklin JM, Platt R, Dreyer NA, London AJ, Simon GE, Watanabe JH, et al. When can nonrandomized studies support valid inference regarding effectiveness or safety of new medical treatments? Clin Pharmacol Ther 2022;111:108—15.
[19] Panagiotou OA, Heller R. Inferential challenges for real-world evidence in the era of routinely collected health data: many researchers, many more hypotheses, a single database. JAMA Oncol 2021;7:1605—7.
[20] Thibault RT, Kovacs M, Hardwicke TE, Sarafoglou A, Ioannidis JPA, Munafò MR. Reducing bias in secondary data analysis via an Explore and Confirm Analysis Workflow (ECAW): a proposal and survey of observational researchers. R Soc Open Sci 2023;10:230568.
[21] Kerr NL. HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev 1998;2:196—217.
[22] Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goettsch W, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Pharmacoepidemiol Drug Saf 2017;26:1033—9.
[23] Ahlbom A. Modern Epidemiology, 4th ed. TL Lash, TJ VanderWeele, S Haneuse, KJ Rothman. Wolters Kluwer, 2021. Eur J Epidemiol 2021;36:767—8.
[24] Desai RJ, Wang SV, Sreedhara SK, Zabotka L, Khosrow-Khavar F, Nelson JC, et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. BMJ 2024;384:e076460.
[25] Naudet F, Patel CJ, DeVito NJ, Le Goff G, Cristea IA, Braillon A, et al. Improving the transparency and reliability of observational studies through registration. BMJ 2024;384:e076123.
[26] Huebner M, Vach W, Le Cessie S, Schmidt CO, Lusa L. Topic Group ''Initial Data Analysis'' of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org). Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Methodol 2020;20:61.
[27] Dal-Ré R, Ioannidis JP, Bracken MB, Buffler PA, Chan AW, Franco EL, et al. Making prospective registration of observational research a reality. Sci Transl Med 2014;6, 224cm1.
[28] FDA/CDER CBER. OCE. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products; 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory#:~:text=Pursuant%20to%20this%20section,%20FDA%20created%20a%20framework%20for%20a. [Accessed 7 October 2024].
[29] Hernán MA, Robins JM. Causal inference. What if; 2020. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. [Accessed 7 October 2024].
[30] Hernán MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004;58:265—71.
[31] Hernán MA. Methods of public health research - strengthening causal inference from observational data. New Engl J Med 2021;385:1345—8.
[32] Chatton A, Rohrer JM. The causal cookbook: recipes for propensity scores, g-computation, and doubly robust standardization. Adv Methods Pract Psychol Sci 2024;7(1), http://dx.doi.org/10.1177/25152459241236149. https://journals.sagepub.com/doi/epub/10.1177/25152459241236149. [Accessed 7 October 2024].
[33] Tennant PWG, Arnold KF, Ellison GTH, Gilthorpe MS. Analyses of 'change scores' do not estimate causal effects in observational data. Int J Epidemiol 2022;51:1604—15.
[34] Naimi AI, Whitcomb BW. Defining and identifying average treatment effects. Am J Epidemiol 2023;192:685—7.
[35] Goetghebeur E, Le Cessie S, Stavola de B, Moodie EE, Waernbaum I, ''on behalf of'' the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. Stat Med 2020;39:4922—48.

[36] Greifer N, Stuart EA. Choosing the causal estimand for propensity score analysis of observational studies. arXiv 2021, http://dx.doi.org/10.48550/arXiv.2106.10577, 2106.10577 [Accessed 7 October 2024].
[37] European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials.; 2020. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf#:~:text=ICH%20E9%20(R1)%20addendum%20on%20estimands%20and%20sensitivity%20analysis%20in. [Accessed 7 October 2024].
[38] Dang LE, Gruber S, Lee H, Dahabreh IJ, Stuart EA, Williamson BD, et al. A causal roadmap for generating high-quality real-world evidence. J Clin Trans Sci 2023;7:e212.
[39] Lipsky AM, Greenland S. Causal directed acyclic graphs. JAMA 2022;327:1083—4.
[40] Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB. Introduction to causal diagrams for confounder selection. Respirology 2014;19:303—11.
[41] Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol 2022;142:264—7.
[42] van Zwieten A, Tennant PWG, Kelly-Irving M, Blyth FM, Teixeira-Pinto A, Khalatbari-Soltani S, et al. Avoiding overadjustment bias in social epidemiology through appropriate covariate selection: a primer. J Clin Epidemiol 2022;149:127—36.
[43] Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun 2020;11:5749.
[44] Piccininni M, Stensrud MJ. Using negative control populations to assess unmeasured confounding and direct effects. Epidemiology 2024;35:313—9.
[45] Groenwold RHH. Falsification end points for observational studies. JAMA 2013;309:1769—70.
[46] Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21:383—8.
[47] Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? JAMA 2013;309:241—2.
[48] Brown JP, Hunnicutt JN, Ali MS, Bhaskaran K, Cole A, Langan SM, et al. Quantifying possible bias in clinical and epidemiological studies with quantitative bias analysis: common approaches and limitations. BMJ 2024;385:e076365.
[49] Haneuse S, VanderWeele TJ, Arterburn D. Using the e-value to assess the potential effect of unmeasured confounding in observational studies. JAMA 2019;321:602—3.
[50] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the e-value. Ann Intern Med 2017;167:268—74.
[51] Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355, i4919.
[52] Fox MP, MacLehose RF, Lash TL. Applying quantitative bias analysis to epidemiologic data. 2nd ed. Springer; 2021. ISBN: 978-3-030-82672-7 (483 pp.).
[53] Zhang H, Clark AS, Hubbard RA. A quantitative bias analysis approach to informative presence bias in electronic health records. Epidemiology 2024;35:349—58.
[54] Dahabreh IJ, Robins JM, Hernán MA. Benchmarking observational methods by comparing randomized trials and their emulations. Epidemiology 2020;31:614—9.
[55] Matthews AA, Dahebreh IJ, MacDonald CJ, Lindahl B, Hofmann R, Erlinge D, et al. Prospective benchmarking of an observational analysis in the SWEDEHEART registry against the REDUCE-AMI randomized trial. Eur J Epidemiol 2024;39:349—61.
[56] Wing K, Williamson E, Carpenter JR, Wise L, Schneeweiss S, Smeeth L, et al. Medications for chronic obstructive pulmonary disease: a historical non-interventional cohort study with validation against RCT results. Health Technol Assess 2021;25:1—70.
[57] Hernán MA, Wang W, Leaf DE. Target trial emulation: a framework for causal inference from observational data. JAMA 2022;328:2446—7.
[58] Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. BMJ 2022;378:e071108.
[59] Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016;79:70—5.
[60] D'Andrea E, Schneeweiss S, Franklin JM, Kim SC, Glynn RJ, Lee SB, et al. Efficacy versus effectiveness: The HORIZON pivotal fracture trial and its emulation in claims data. Arthritis Rheumatol 2024, http://dx.doi.org/10.1002/art.42968. https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/art.42968. [Accessed 7 October 2024].
[61] FDA. Guidance document. Demonstrating substantial evidence of effectiveness for human drug and biological products. Draft guidance for industry.; 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/demonstrating-substantial-evidence-effectiveness-human-drug-and-biological-products. [Accessed 7 October 2024].
[62] Wang SV, Schneeweiss S, Franklin JM, Desai RJ, Feldman W, et al., RCT-DUPLICATE Initiative. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. JAMA 2023;329:1376—85.
[63] Heyard R, Held L, Schneeweiss S, Wang SV. Design differences and variation in results between randomised trials and non-randomised emulations: meta-analysis of RCT-duplicate data. BMJMed 2024;3:e000709.
[64] Zhao SS, Lyu H, Solomon DH, Yoshida K. Improving rheumatoid arthritis comparative effectiveness research through causal inference principles: systematic review using a target trial emulation framework. Ann Rheum Dis 2020;79:883—90.
[65] Zuo H, Yu L, Campbell SM, Yamamoto SS, Yuan Y. The implementation of target trial emulation for causal inference: a scoping review. J Clin Epidemiol 2023;162:29—37.
[66] Hiemstra B, Keus F, Wetterslev J, Gluud C, van der Horst ICC. DEBATE-statistical analysis plans for observational studies. BMC Med Res Methodol 2019;19:233.
[67] Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ 2018;363, k3532.
[68] Wang SV, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin JA, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. BMJ 2021;372:m4856.
[69] Orsini LS, Monz B, Mullins CD, Van Brunt D, Daniel G, Eichler HG, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. Pharmacoepidemiol Drug Saf 2020;29:1504—13.